

BOLT : Efficient Betweenness Ordering in almost Linear Time

Manas Agarwal², Rishi Ranjan Singh¹, Shubham Chaudhary², S.R.S. Iyengar¹

¹ *Department of Computer Science and Engineering, Indian Institute of Technology Ropar
Nangal Road, Rupnagar, Punjab, India*

² *Department of Mathematics, Indian Institute of Technology Roorkee
Roorkee, Uttarakhand, India*

¹{rishirs, sudarshan}@iitrpr.ac.in

²{manasuma, shubhuma}@iitr.ac.in

Abstract—Centrality measures, erstwhile popular amongst the sociologists and psychologists, have seen wide and increasing applications across several disciplines of late. In conjunction with the big data problems there came the need to analyze big networks and in this connection, centrality measures became of great interest to the community of mathematicians, computer scientists and physicists. While it is an important question to ask how one can rank vertices based on their importance in a network, there hasn't been a commonly accepted definition, mainly due to the subjectivity of the term "importance". Amongst a plethora of application specific definitions available in the literature to rank the vertices, closeness centrality, betweenness centrality and eigenvector centrality (page-rank) have been the most important and widely applied ones. In the current paper, we formulate a method to determine the betweenness ordering of k vertices without exactly computing their betweenness indices - which is a daunting task for networks of large size. The method results very efficient ordering even when runs for linear time in the number of edges. We apply our approach to find the betweenness ordering of k vertices in several synthetic and real world graphs. We compare our method with the available techniques in the literature and show that our method fares several times better than the currently known techniques. We further show that the accuracy of our algorithm gets better with the increase in size and density of the network.

Index Terms—Centrality Ordering, Betweenness Centrality, Betweenness Ordering, Nonuniform Sampling, Approximation.

1 INTRODUCTION

Centrality of a vertex in a network is the quantification of the intuitive notion of importance of a node in a network. Centrality measures have been extensively used in the analysis of large data available from real world networks in the recent times. Centrality indices, also referred as structural indices, are real valued functions that remain invariant under isomorphic transformation of graphs [1]. Different centrality measures are coined in the literature for application specific reasons. For a detailed study of centrality indices and their applications, one can refer to the books by Newman [2], Jackson [3] and Brandes and Erlebach [1]. Real-world networks are generally very large in size, dynamic in nature and keep changing at a very high rate. In such networks, comparing centrality scores of two nodes is of great importance, consider for example, a production company is finalizing a new brand ambassador for their organization and has two options to choose from. To evaluate which one is better, one might need to compare the importance (in this case popularity) of the two candidate actors in a given social network where the number of nodes are in the order of thousands. Consider two research-papers in the citation network [4] which is of the order of a few million nodes, how can one find which paper is more central than the other one? For example, if one were to compute betweenness centrality in this case, even with the adoption of the best known algorithm, it is a time consuming task for large sized networks. We ask this question *Is there a method*

to compute the centrality ordering of two nodes and declare which one is more central than the other, without actually computing their exact centrality values. More formally, given two nodes u and v in a graph G with centrality values $C(u)$ and $C(v)$ with $C(u) > C(v)$ thus making u superior in rank than v , can one get to know which node is of superior rank over the other without computing its centrality values. We call this problem the *centrality-ordering* problem.

We illustrate centrality ordering problem in the context of a centrality measure called the *eccentricity* measure and give a simple approximation approach for eccentricity ordering.

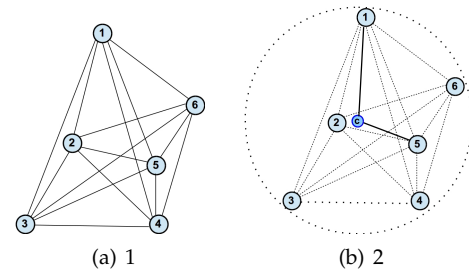


Fig. 1. Eccentricity ordering in 2-D Euclidean plane

Eccentricity of a node v [5] in a connected graph G is defined as the shortest distance to the farthest node

from v in G . Center of a graph which is a solution to the *facility location problems*, is calculated by picking the nodes with least eccentricity. Finding eccentricity of all nodes is as expensive as finding closeness, betweenness or stress centrality for all nodes in respect of time. Given a graph in the two dimensional euclidean space, if we were to solve the *eccentricity ordering problem* of two nodes in that graph without computing the eccentricities, we would go about the following way: drawing a minimum circle (Disk) covering all the nodes is a very well known problem called *smallest-circle problem* or *minimum covering circle problem*. A linear ($O(n)$) time randomized algorithm by Welzl [6] can find the smallest circle covering n points on a 2-D euclidean plane. Once we find the smallest circle, an approximate solution to the eccentricity comparison problem is to compare the distance from the center of smallest circle to the nodes. If the nodes are evenly distributed in the smallest circle, then the node closer to the center of that smallest circle is likely to have smaller eccentricity and vice versa. We just noted that we could solve the eccentricity ordering problem in linear time as opposed to finding it the conventional way by considering all possible distances from the given vertex to all other vertices.

In general, there are three possible ways to solve the centrality-ordering problem if we allow exact computation of centrality scores:

- 1) Compute exactly the centrality scores of both nodes and compare the scores to get the ordering.
- 2) Approximate the centrality scores of both nodes efficiently and compare the scores to get the approximate ordering.
- 3) Directly compute the exact or approximate centrality-ordering exploiting some structural property of the given network.

The reason, we exclude the first type of solution for ordering problem is summarized below. This trivial method for exact centrality ordering calculates the centrality score of both the nodes and then compares the values in order to answer which one is more important. There are two reasons why the current state of the art algorithms for exact calculation of the centrality measures are not time efficient. Firstly because of the large size and the dynamic nature of networks. In large dynamic networks, we have to recompute the centrality scores each time the network changes, which is evidently expensive. Secondly because of the global characteristics of some centrality measures. For example, closeness centrality and degree centrality computation of a single node takes very less time as compared to the computation of the respective centrality measures for all the nodes in network. But unlike degree and closeness centralities, computing betweenness centrality of a node is conjectured to be as expensive as computing it for all the nodes in any network [7].

To reduce the time required for ordering, the second method efficiently estimates the individual centrality scores and finds the correct ordering with a high probability. The third and the last type of solution is, where without computing or estimating the centrality scores, we order the nodes based on some structural properties in the network. Such a

type of solution is given in previous paragraph for a special case of eccentricity-ordering.

Since, computing the betweenness centrality of one node is equivalent to computing the betweenness centrality of all nodes according to the currently known deterministic algorithms, we are motivated to address the problem of betweenness ordering of two vertices. First we give an algorithm to find the betweenness ordering of two vertices (here onwards called the betweenness-ordering-problem). Then, we extend the algorithm for comparing k vertices where $2 < k \ll n$ and n is the total number of nodes. In this paper we propose a second type of solution for betweenness-ordering problem based on a nonuniform sampling based efficient estimation. We explain in detail the estimation algorithm given in the shorter version [8] on the paper and then discuss the betweenness-ordering results based on it.

Betweenness centrality was proposed by Freeman [9] and Anthonisse [10] independently. Betweenness centrality of a node v is defined as the relative fraction of shortest paths passing through v . It is calculated as $BC(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$, where σ_{st} is the total number of shortest paths from vertex s to vertex t and $\sigma_{st}(v)$ is the total number of shortest paths from vertex s to vertex t passing through vertex v . Unlike degree centrality, betweenness centrality covers more global characteristics and unlike closeness centrality works even on disconnected networks. Betweenness centrality has found many important applications in diverse fields. It has been used in the biological networks [11], protein-protein interaction (PPI) networks [12], analyzing communication system networks [13], identifying critical nodes in the electrical & electronic systems (EES systems like Electronic Control Units used in vehicles) [14], analyzing supply chain networks [15], identifying bottleneck in supply chain networks [16], planning a better public transit system networks; for example metro networks [17], measuring load at a node in gas pipeline network [18], waste-water disposal system networks etc.

We present a real world example that better motivates the betweenness-ordering problem. *Cascading failure* in networks is a phenomenon where a node's failure triggers the failure of successive nodes. After a node fails, the load of that node gets distributed among the neighbour nodes. This event may overload few neighbour nodes, which, in turn, may overload their neighbour nodes and goes on and on. This process of successive failure of nodes may cause the whole network to fail. Blackouts (power outage) are the most common outcomes of cascade failure. A list of major blackouts and other examples of cascading failure can be found at the wikipedia web page¹. The most recent blackout due to cascading failure had happened in India on July 30, 2012. Blackout in Italy on September 28, 2003 resulted in a widespread failure of other dependent systems; railway network, health care systems, financial services, communication networks [19]. The cascading failure must be avoided. Several models have been proposed for cascading failure [20], [21], [22], [23]. Some of them have shown that the failure of a node having high betweenness score may cause greater collapse of the network. Supposedly, two

1. http://en.wikipedia.org/wiki/Cascading_failure

nodes in such a network failed at the same time, one should focus on the recovery/maintenance of the node with higher betweenness. This problem requires a faster betweenness ordering of those two nodes. Let Fig. 2 be a power grid network (a rectilinear grid of dimension 2×2 where the number of nodes are 9 and the number of edges are 12). The betweenness score of each node is written next to it. Now if nodes 2 and 8 fail at the same time and we have a single resource available to recover only one of them, then according to the above theory we should send the resource to node 2.

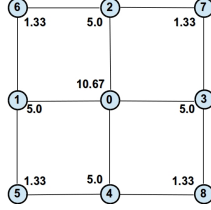


Fig. 2. Example rectilinear grid network of size 2×2

This is a revised and an extended version of the results that appeared in CompleNet 2015 [8]. In this paper, we discuss a novel non-uniform sampling technique which better approximates the optimal sampling explained in [24]. We efficiently estimate the betweenness score of a given node using the proposed sampling model incorporated within the approximation algorithm, given in section 3.

The contribution of this paper are as follows.

- 1) To the best of our knowledge, it is the first study that focuses on the ordering of nodes based on centrality scores.
- 2) We devise an almost linear time approach for betweenness-ordering of two nodes and further generalize it for ordering an arbitrary k nodes. First we discuss a very efficient non-uniform sampling technique to choose the source nodes (also called pivot nodes) for single source shortest path computation. Then we use the model to approximately compute the betweenness score of a given node without computing betweenness of all the nodes in the graph.
- 3) The developed non-uniform node sampling model provides the best approximation of the optimal sampling model given in [24].
- 4) We conduct extensive simulations on real and synthetic networks. The results show that our algorithm and sampling model outperforms most of the sampling based or deterministic approximation algorithms. We measure the performance in ordering using a performance tool defined in the paper and show that the proposed approach is very efficient.

We organize the rest of the paper as follows. In next section, we briefly review the approaches related to the betweenness computation. In section 3, we define basic terms used in the paper and explain the previous concepts, based on which we develop our sampling model. In section 4, we develop our model based on the analysis of random

networks and some observations. All the details about simulations, data sets used in simulations, performance tools used for evaluation and extensive results in the form of plots and tables are compiled in section 5. We discuss the possible future directions of work in section 6. We conclude the paper in section 7.

2 RELATED WORK

Ordering of betweenness centrality can find several applications in various real-world scenarios. To the best of our knowledge, it is the first work which considers and motivates the study on the betweenness ordering of two nodes. Most of the studies done so far consider only computation of betweenness scores or ranking all the nodes based on their betweenness score. We discuss in brief about few of the computation algorithms. All the exact computation algorithms are based on either single source shortest path (SSSP) computation algorithms from all sources or all pair shortest path computation algorithms. The most trivial algorithm is a modified version of the Floyd-Warshall's APSP algorithm [25], [26] to compute the betweenness scores for all nodes [9]. But this takes $O(n^3)$ time where n is the number of nodes. In year 2001, Brandes [27] introduced an algorithm based on Dijkstra's algorithm [28] which computes the exact betweenness score of all nodes in unweighted graphs in $O(mn)$ time, where m is the number of edges.

Due to the size of current real world networks, even the state of art (Brandes') algorithm was very expensive in terms of time. This motivated the researchers to develop faster exact or approximation algorithms. Several exact algorithms for large graphs (Sariyüce et al. [29]) and dynamic graphs (Lee et al. [30], Green et al. [31], Kas et al. [32], Goel et al. [33], Nasre et al. [34]) have been developed. These algorithms improved the computation time experimentally on special type of graphs but in worst case they all were as expensive as Brandes' [27]. Several approximation algorithms were also proposed. These algorithms ran much faster and computed centrality scores close to the exact centrality scores. Two types of approximation algorithms exist. The First type consists of algorithms that focus on computing the approximate betweenness of all nodes together. The second type comprises of algorithms that approximate the betweenness score of a given node. Our goal is also to develop an approximation algorithm, therefore we summarize most of the approximation ideas developed so far for the betweenness computation.

Eppstein and Wang [35] first proposed the idea of sampling to approximately compute centrality indices for which SSSP computation is required from all the nodes. They suggested to compute SSSP from only a few nodes (called pivot) and discussed how to approximate the closeness centrality. Brandes and Pich [36] extended the idea of sampling given by Eppstein and Wang for approximating betweenness centrality. They gave different pivot selection strategies. SSSP from each pivot node were computed to estimate the contribution of each pivot node in the betweenness score of all nodes. By extrapolating the average contribution from pivot nodes, the approximate betweenness centrality was computed.

Bader et al. [37] proposed an adaptive sampling based approximation algorithm to compute the betweenness score of a given node. In his study, uniform probabilities were considered to sample the nodes. The number of sampled nodes were dependent on the importance of the considered node, i.e. for highly central nodes, the algorithm requires to sample less number of nodes as compared to the nodes sampled for less central nodes. They also provided theoretical bound for their approximation algorithm. Geisberger et al. [38] generalized the approach given by Brandes and Pich [36] and observed that the betweenness centrality scores of unimportant (less betweenness central) nodes which are near to pivot nodes, get overestimated. They provided an unbiased betweenness estimator framework which overcomes the observed problem. Gkorou et al. [39] developed two approximation approaches to compute approximate betweenness. Their first approach was for the dynamic networks and was based on the observation that very highly central nodes remain almost invariant over dynamic operations. The second approach was for large networks and considered only k -length shortest paths for the computation of approximate betweenness score. Their algorithms did not perform well on random graphs. Riondato and Kornaropoulos recently [40] developed two randomized algorithms to approximate betweenness score based on sampling of shortest paths and analyzed theoretically. The first algorithm approximates the betweenness score for all the nodes and the second algorithm approximates the betweenness score for top- k nodes.

Recently, Chehreghani [24] proposed a new idea of approximating the betweenness score of a given node. He used non-uniform sampling and then unlike [36], [38], he scaled the contributions from sampled nodes with respect to the probabilities. Finally, he averaged all the scaled values to achieve the approximate score. He used a very trivial model for generating the non-uniform probabilities but has not given any theoretical proof for the formula used in that model.

3 PRELIMINARY

In this section, we introduce some basic terms related to the betweenness centrality which have been used throughout the paper. We also discuss the previous concepts that motivated our sampling technique.

3.1 Terminology

We use following terms interchangeably; node or vertex and graph or network. For simplicity, we consider only unweighted undirected graphs until mentioned explicitly. All the concepts discussed in this paper can be easily generalized for weighted or directed graphs. Given a graph $G = (V, E)$, V is the set of nodes with $|V| = n$ and E is the set of edges with $|E| = m$. A (simple) *path* is a sequence of edges connecting a sequence of vertices without any repetition of vertices. The *length* of a path is the number of edges in the path. *Shortest paths* between two vertices are the smallest length paths between them. *Distance* between two nodes i and j , $d(i, j)$, is the length of shortest path between i and j .

Algorithm 1 : Approximation algorithm to compute betweenness score of a given node v [24].

Estimate(G, P, v)

- 1: **Input.** Graph G , probabilities $P = \{p_1, p_2, \dots, p_n\}$, node v .
 - 2: $BC(v) = 0$.
 - 3: **for** $i=1$ to T **do**
 - 4: Select a node i with probability p_i .
 - 5: Compute $\delta_{i\bullet}(v)$ in the BFT_i using equation (1).
 - 6: $BC(v) \leftarrow BC(v) + \frac{\delta_{i\bullet}(v)}{p_i}$.
 - 7: **end for**
 - 8: $BC[v] \leftarrow BC(v)/T$.
 - 9: **Return.** $BC(v)$.
-

Let σ_{st} be the number of shortest paths between s and t , for $s, t \in V$. Let $\sigma_{st}(v)$ be the number of shortest paths between s and t passing through v , for $v \in V$. Betweenness centrality score of a node $v \in V$ is calculated as

$$BC(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}.$$

Pair dependency of a pair of vertices (s, t) on a vertex v is defined as: $\delta_{st}(v) = \frac{\sigma_{st}(v)}{\sigma_{st}}$. Betweenness centrality of a vertex v can be defined in terms of pair dependency as

$$BC(v) = \sum_{s \neq v \neq t \in V} \delta_{st}(v).$$

Let BFT_r denotes the breadth-first traversal (BFT) of the graph rooted on vertex r . In BFT_r , we assume that r is at level 0 and the next levels are labelled by natural numbers in an increasing order. *Dependency* of a vertex s on a vertex v is defined as: $\delta_{s\bullet}(v) = \sum_{t \in V \setminus \{s, v\}} \delta_{st}(v)$. Let us define a set $P^s(w) = \{v : v \in V, w \text{ is a successor of } v \text{ in } BFT_s\}$. Brandes [27] proved that:

$$\delta_{s\bullet}(v) = \sum_{w: v \in P^s(w)} \frac{\sigma_{sv}}{\sigma_{sw}} (1 + \delta_{s\bullet}(w)). \quad (1)$$

3.2 A Betweenness Approximation Technique Based on Non-uniform Sampling

In this section we will briefly describe the recent work of Chehreghani [24] that provides motivation for our work. He gave an approximation algorithm to compute betweenness score of a given node v . The algorithm is summarized as Algorithm 1. For a given node v , the algorithm takes the sampling probabilities as input and outputs the approximate betweenness score of node v . Step 2 initializes the betweenness score to 0. The algorithm estimates the betweenness score of node v , T number of times and takes the average of all T estimations. In each iteration of the algorithm, it samples a pivot node and computes the dependency of the pivot node on node v using a single iteration of Brandes' algorithm [27]. Then it estimates the betweenness score of node v by, dividing (scaling) the computed dependency by the sampling probability of that pivot node. He has motivated his paper with the idea of optimal sampling that is stated in the following theorem.

Theorem 1. [24] Let the sampling probability assigned to each node i be

$$p_i = \frac{\delta_{i\bullet}(v)}{\sum_{j=1}^n \delta_{j\bullet}(v)}$$

then, betweenness score of node v can be exactly calculated in $O(m)$ time using single iteration of Algorithm 1.

We refer the probability defined in Theorem 1 as *optimal probability* and call a model *optimal model (OPT)* if it can generate optimal probabilities. Calculating optimal probabilities is as expensive as computing exact betweenness using Brandes' algorithm [27]. Thus, a model was desired that can efficiently estimate sampling probabilities close to the optimal. Chehreghani noted that any such model should satisfy at least the following relation for most of the vertex pairs (i, j) :

$$p_i < p_j \iff \delta_{i\bullet}(v) < \delta_{j\bullet}(v) \quad (2)$$

Chehreghani has given a simple distance based model (DBM) [24] to generate the sampling probabilities. He proposed to take the probabilities as the normalized value of the inverse of distance from node v to node i , $p_i \propto \frac{1}{d(v,i)}$. He has shown experimentally that his nonuniform sampling technique reduces the error in the computation of betweenness score as compared to uniform sampling technique ([36], [37]). But, he was unable to provide a theoretical derivation for DBM. In DBM, many of the nodes j with $\delta_{j\bullet}(v) = 0$ get same probabilities as nodes i with $\delta_{i\bullet}(v) \neq 0$ because of being at the same level in BFT_v . We next propose a new probability estimation model for nodes that efficiently approximates the optimal probabilities and outperforms DBM.

4 A NEW NON-UNIFORM SAMPLING MODEL

We consider the problem of ordering two nodes based on betweenness centrality. To order the two nodes based on their betweenness scores, we first approximately compute their individual scores and then compare. Our main problem reduces to a sub-problem which requires computing a very efficient approximation of the betweenness score of a given node. In this section, we discuss a model which generates non-uniform probabilities for sampling the nodes. This model can be incorporated with Algorithm 1 to solve the above sub-problem. Our model is based on the inverse of degree and an exponential function in the power of distance, thus we refer it as EDDBM (exponential in distance and inverse of degree based model). The developed model reduces the average error by generating probabilities very close to the optimal probabilities. We try assigning larger probability values to the vertices contributing more to the betweenness of a given node v and smaller to those which contribute less. We perform the analysis on the random graphs to establish the relation between probabilities and distance. Then, on the basis of few observations, we achieve relation between the probability and degree. At last we describe the steps to generate the probabilities by our model.

4.1 Analysis of Random Graphs

Let G be a random graph that is generated based on the $G(n, p)$ model given by Erdos and Renyi [41]. We are given a vertex v to compute its betweenness score. We first analyze how the dependency of a node i on the node v ,

$\delta_{i\bullet}(v)$ varies when v lies on different levels in BFT_i . This will help us to establish a relation between $\delta_{i\bullet}(v)$ and the distance between i and v . For this, first we need to compute the expected number of nodes at any level m of a BFS traversal. Wang [42] gave a complex approach to estimate the number of nodes at any level in BFS traversal on various types of graphs based on generating functions and degree distribution. We give a simple approach to estimate the number of nodes at a level in BFS traversal in random $G(n, p)$ graphs. Let λ be the average degree of the given graph and let p be the probability of an edge's existence. The first lemma approximately estimates the number of nodes at a given level in a BFS traversal by a recurrence relation.

Lemma 1. Let α_j be the number of nodes at level j in the BFS_i . Then the number of nodes at level $m + 1$, α_{m+1} can be given as:

$$\alpha_{m+1} \approx np(1 - \frac{\sum_{j=0}^m \alpha_j}{n})\alpha_m. \quad (3)$$

Proof. Van Der Hofstad [43] explained the BFS traversal as Exploration Technique (ET) in random graphs. In this technique all the vertices are initially inactive except i (the root node on which ET has to be applied). The vertex w is chosen which was discovered first among the current active vertices, and its neighbourhood is explored for inactive vertices. All the inactive vertices found are made active. Node w is made inactive and is labelled as processed. In the paper we refer exploring the neighbourhood of t^{th} vertex as t^{th} exploration.

Let S_t be the number of active vertices after t^{th} exploration, and X_t be the number of vertices discovered (converted from inactive to active) in t^{th} exploration. Then, following relation holds for any iteration (exploration) t :

$$S_t = S_{t-1} + X_t - 1 \quad (4)$$

After $t - 1$ explorations, we are left with $n - (t - 1) - S_{t-1}$ vertices ($t - 1$: processed vertices, S_{t-1} : active vertices). If p is the probability of existence of an edge between any two nodes in the graph then we have:

$$X_t \sim \text{Bin}(n - (t - 1) - S_{t-1}, p) \quad (5)$$

Next, we state a very well known binomial relation in mathematics as lemma. Then we will use it to compute the expected number of nodes at any level of BFS traversal.

Lemma 2. If $X \sim \text{Bin}(n, p)$ then, $\Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$. The expected value of X , $E[X]$ is:

$$E[X] = \sum_{k=1}^n k \binom{n}{k} p^k (1 - p)^{n-k} = np.$$

Using lemma 2 we can write the expected value of X_t as:

$$E[X_t] = (n - (t - 1) - S_{t-1})p \quad (6)$$

Equation (6) can be used in the following way to calculate the expected number of nodes at a BFS level. Initially,

there is a single (source) node as an active node i.e. $S_0 = 1$. The expected number of nodes discovered in the first exploration will be:

$$E[X_1] = (n - 1)p$$

After first exploration, the total number of active nodes will be $S_1 = (n - 1)p$. The expected number of nodes discovered in the second exploration will be $(n - 1 - (n - 1)p)p$ or

$$E[X_2] = (n - 1)(1 - p)p.$$

Similarly, we can calculate following values: $S_2 = (n - 1)(2 - p)p - 1$, $E[X_3] = (n - 1)(1 - p)^2p$, $E[X_4] = (n - 1)(1 - p)^3p$. In the above manner we can calculate the number of active nodes before each exploration and the expected number of nodes discovered in each exploration.

Let α_j be the expected number of nodes at level j . We have $\alpha_0 = 1$ and $\alpha_1 = (n - 1)p$. Then by using equation (6) we can calculate α_2 as:

$$\alpha_2 = \sum_{k=1}^{(n-1)p} (n-1)(1-p)^k p = (n-1)(1-p)[1 - (1-p)^{(n-1)p}] \quad (7)$$

Now, we can derive the formula for the general case. Let us assume that $m - 1$ levels have been explored, i.e. the nodes of level m have been discovered. Now we have to explore the nodes of level m . Exploring the first vertex of level m discovers $(n - \sum_{j=0}^m \alpha_j)p$ nodes for level $m + 1$. Exploration of the next vertex discovers $[(n - \sum_{j=0}^m \alpha_j) - (n - \sum_{j=0}^m \alpha_j)p]p = (n - \sum_{j=0}^m \alpha_j)(1 - p)p$ nodes and so on. So the expected number of nodes at level $m + 1$ will be:

$$\alpha_{m+1} = \sum_{k=0}^{\alpha_m-1} (n - \sum_{j=0}^m \alpha_j)(1 - p)^k p$$

or

$$\alpha_{m+1} = (n - \sum_{j=0}^m \alpha_j)[1 - (1 - p)^{\alpha_m}] \quad (8)$$

Most of the real world graphs are sparse. Hence, without the loss of generality, we can assume that $p \ll 1$. Applying Binomial expansion and neglecting higher order terms of p , we can rewrite equation (8) as:

$$\alpha_{m+1} \approx (n - \sum_{j=0}^m \alpha_j)\alpha_m p = n(1 - \frac{\sum_{j=0}^m \alpha_j}{n})\alpha_m p.$$

□

Equation (3) is a recurrence relation to estimate the number of nodes at some level $m + 1$. Using Lemma 1, we can estimate the ratio between the expected number of nodes at two consecutive levels. The ratio is derived as follows.

In random graphs, the average degree λ is equal to $(n - 1)p$, where n is the number of nodes and p is the existential probability of an edge. λ can be approximated as np . If we denote $(1 - \frac{\sum_{j=0}^m \alpha_j}{n})$ as c_{m+1} (the fraction of

nodes below level m), then we can rewrite equation (3) as $\alpha_{m+1} \approx c_{m+1}\lambda\alpha_m$ or

$$\frac{\alpha_{m+1}}{\alpha_m} \approx c_{m+1}\lambda \quad (9)$$

where $c_{m+1} \in [0, 1)$.

Based on equation (9), we derive the formula to calculate the expected dependency of a node i on node v , $E[\delta_{i\bullet}(v)]$ in next lemma. Then, we establish the ratio between the expected dependency of root node i on two nodes at consecutive levels in Theorem 2.

Lemma 3. Let v be a node at m th level in the BFS S_i . Then the expected dependency of node i on node v can be given as

$$E[\delta_{i\bullet}(v)] = (\frac{\alpha_{m-1}}{\alpha_{m-2}})(1 + c_m\lambda) = c_{m-1}\lambda(1 + c_m\lambda). \quad (10)$$

Proof. Let the BFS traversal rooted at i consist of $m + 1$ levels. If v is at the last level (m), then $\delta_{i\bullet}(v) = 0$. Now, if v lies at level $m - 1$, then we can compute the expected dependency ($E[\delta_{i\bullet}(v)]$) as follows. Let A_{m-1} be the expected number of paths of length $m - 1$ from i to any vertex at level $m - 1$. Bauckhage et al. [44] gave following expression for A_{m-1} :

$$A_{m-1} = n^{m-2}\pi^{m-1} \quad (11)$$

where $\pi = \frac{n-1}{n}$. It is simple to observe that any node w at level m has $\alpha_{m-1} \cdot p$ parents (nodes at level $m - 1$ which are connected to w by a direct edge), so the expected number of shortest paths from i to w will be $A_{m-1} \cdot \alpha_{m-1} \cdot p$. Node v lies only on A_{m-1} paths out of those shortest paths. From equation (1), it is easy to observe that the expected partial dependency from node i to node w on node v is $E[\delta_{iw}(v)] = \frac{1}{\alpha_{m-1}p}$. Node v has $\alpha_m \cdot p$ children similar to w . Therefore the expected dependency of node i on node v is $E[\delta_{i\bullet}(v)] = \frac{\alpha_m}{\alpha_{m-1}}$ or using equation (9) we can rewrite:

$$E[\delta_{i\bullet}(v)] = c_m\lambda.$$

Similarly, if v lies at level $m - 2$, then the expected dependency of node i on node v can be given as:

$$E[\delta_{i\bullet}(v)] = (\frac{\alpha_{m-1}}{\alpha_{m-2}})(1 + c_m\lambda) = c_{m-1}\lambda(1 + c_m\lambda).$$

□

Now, we can give the theorem stating the ratio between dependencies of root node on two nodes positioned at two consecutive levels.

Theorem 2. Let l be the last level in BST $_i$. Let $\delta_{i\bullet}(v_{l-k})$ be the dependency of node i at a node v_{l-k} at level $l - k$ and let $\delta_{i\bullet}(v_{l-k+1})$ be the dependency of node i at a node v_{l-k+1} at level $l - k + 1$. Then we have

$$\frac{\delta_{i\bullet}(v_{l-k})}{\delta_{i\bullet}(v_{l-k+1})} = c_{l-k+1}(\frac{1}{\phi} + \lambda) \quad (12)$$

where $\phi = (c_{l-k+2})(1 + c_{l-k+3}\lambda(1 + c_{l-k+4}\lambda(1 + c_{l-k+5}\lambda(1 + \dots(1 + c_l\lambda)) \dots))$.

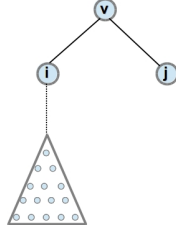


Fig. 3. An example BST_v

Proof. The ratio of expected dependencies of node i on v , when v lies at level $l - 2$, ($E[\delta_{i\bullet}(v_{l-2})]$) to when v lies in level $l - 1$, ($E[\delta_{i\bullet}(v_{l-1})]$) is

$$\frac{\delta_{i\bullet}(v_{l-2})}{\delta_{i\bullet}(v_{l-1})} = \frac{c_{l-1}}{c_l} (1 + c_l \lambda). \quad (13)$$

In general, the ratio of the expected dependencies for two successive levels $l - k$ and $l - k + 1$ can be given as equation (12). \square

It is simple to observe that c_m decreases continuously as m increases. So as v becomes closer to i , $E[\delta_{i\bullet}(v)]$ increases steeply, proportional to the average degree λ . Therefore, on the basis of Theorem 2, we can assign a probability p_i to the node i as in the following theorem.

Theorem 3. Suppose, we have to compute the betweenness score of node v . Then the sampling probability assigned to node i is :

$$p_i \propto (\lambda)^{-d(i,v)} \quad (14)$$

where $d(i, v)$ is the distance between v and i .

4.2 Further Tweak

In this section, we discuss some of the observations and propose some possible solutions to tackle an observed problem.

In BFT_v , nodes at same level are called siblings. We define successors of a node j in BFT_v , $Succ_v(j)$, as the set of nodes to which at least one shortest path from v passes through j . Similarly, we define predecessors of a node j in BFT_v , $Pred_v(j)$, as the set of predecessors. Let $Reach_v^j$ be the set of nodes that are at most as far as v from j .

Observation 1. In the BFS tree rooted at the given node v , siblings get equal probabilities by equation (14) but might not contribute equally in the betweenness of v .

For example, Fig. 3 shows a BFS traversal rooted at node v of some graph. We can notice that node i and node j are at the same level. So, according to Equation 14, equal probabilities will be assigned to both nodes. But $\delta_{i\bullet}(v) = 1$ and $\delta_{j\bullet}(v) = n - 2 \gg 1$. Thus we need to tweak the formula to resolve this problem.

Observation 2. In BST_i , no node from $Succ_v(i) \cup Pred_v(i) \cup Reach_i^v$ will contribute in $\delta_{i\bullet}(v)$.

Observation 2 infers that in BFT_v , a node i with larger number of successors will contribute ($\delta_{i\bullet}(v)$) lesser i.e. the relation can be assumed as $\delta_{i\bullet}(v) \propto$

$\frac{1}{|Succ_v(i) \cup Pred_v(i) \cup Reach_i^v|}$. Then based on the assumption, the probability assigned to node i should also satisfy the following relation:

$$p_i \propto \frac{1}{|Succ_v(i) \cup Pred_v(i) \cup Reach_i^v|}.$$

Maintaining the sets $Succ_v(i)$, $Pred_v(i)$, and $Reach_i^v$ for each node in BFT_v (graph) can not be achieved in linear time. At place of $Succ_v(i) \cup Pred_v(i) \cup Reach_i^v$, we use degree of node i . One reason for using degree is that it can be linearly computed and most of the time, it is the best predictor for the number of successors, predecessors in random graphs. Another reason is the high correlation between the Betweenness centrality and the Degree centrality [45], [46].

Thus, to overcome the problem stated in Observation 1, we also include the following relation:

$$p_i \propto \frac{1}{deg(i)} \quad (15)$$

where $deg(i)$ is the degree on node i in the given graph. Thus the final relation can be written as

$$p_i \propto \frac{(\lambda)^{-d(v,i)}}{deg(i)}$$

We use the distance as inverse power in the exponential function over average degree, and inverse of degree for modelling the probability generation. Thus we name this model EDDBM (exponential in distance and degree based model). Next, we discuss the steps to generate the probabilities according to EDDBM.

4.3 EDDBM

We generate the probabilities as following. First, we generate the probabilities on the basis of distance relation given in Equation 14. Each node i at level d in the BFT_v will get following probability values:

$$p^d = \frac{(\lambda)^{-d}}{\sum_{j \in V \setminus \{v\}} (\lambda)^{-d}}$$

Let V_d be the set of nodes at level d in the BFT_v and $|V_d|$ denotes the number of nodes in set V_d . Then to resolve the problem stated in Observation 1 to best extent, at each level d , we further tweak the formula on the basis of Equation 15 and get the assigned probability to node i at d^{th} level as:

$$p_i = \frac{p^d |V_d| \cdot deg(i)^{-1}}{\sum_{j \in V_d} deg(j)^{-1}}. \quad (16)$$

5 ALMOST LINEAR TIME ALGORITHM FOR BETWEENNESS - ORDERING

In this section, we discuss our almost linear time approach for solving the betweenness-ordering problem based on the new non-uniform based sampling EDDBM. Then we discuss betweenness-ordering problem on k nodes, we refer it as k -betweenness-ordering.

Algorithm 2 : Betweenness-Ordering algorithm.*Betweenness_Ordering*(G, u, v)

- 1: **Input.** Graph G , node u and node v .
- 2: Generate P_u , set of probabilities for each node based on EDDBM model in BFT_u .
- 3: Estimate the betweenness score of node u , $B'(u) = Estimate(G, P_u, u)$.
- 4: Generate P_v , set of probabilities for each node based on EDDBM model in BFT_v .
- 5: Estimate the betweenness score of node v , $B'(v) = Estimate(G, P_v, v)$.
- 6: **Return.** The result of comparison between $B'(u)$ and $B'(v)$.

5.1 Betweenness-Ordering : Ordering 2 nodes

Given a graph G , this algorithm orders two nodes u and v by first efficiently estimating their betweenness scores and then comparing the scores. The algorithm is summarized as Algorithm 2.

Step 2 and step 3 estimates betweenness score of node u . Step 4 and step 5 estimates the betweenness score of node v . Step 2 (step 4) generates non-uniform sampling probabilities using the EDDBM model (equation 16) in relation to node u (v). Step 3 (step 5) estimates betweenness score of node u (v) by passing the generated probabilities to Algorithm 1.

5.2 k -Betweenness-Ordering : Ordering k nodes

Given a graph G , this algorithm orders k nodes by efficiently estimating their betweenness scores and then sorting nodes based on their estimated betweenness scores. The algorithm is summarized as Algorithm 3.

Algorithm 3 : k -Betweenness-Ordering algorithm. k — *Betweenness_Ordering*(G, U)

- 1: **Input.** Graph G , A set $U = v_1, v_2, \dots, v_k$ of k nodes.
- 2: **for** $i=1$ to k **do**
- 3: Set $B'(v_i) = 0$.
- 4: Generate P_{v_i} , set of probabilities for each node based on EDDBM model in BFT_{v_i} .
- 5: Estimate the betweenness score of node v_i , $B'(v_i) = Estimate(G, P_{v_i}, v_i)$.
- 6: **end for**
- 7: Sort the nodes using Merge sorting algorithm [47] based on their betweenness scores.
- 8: **Return.** The nodes in sorted order.

Steps 3-5 estimate betweenness score of node v_i . Step 4 generates non-uniform sampling probabilities using the EDDBM model (equation 16) in relation to node v_i . Step 5 estimates betweenness score of node v_i by passing the generated probabilities to Algorithm 1. Step 7 sorts nodes based on their estimated betweenness scores using merge sort algorithm [47].

5.3 Computation Time

The time complexity of procedure *Estimate*(G, P, v) (Algorithm 1) is $O(Tm)$ [24], where m is the number of edges in

the graph G . It is due to T iterations of breadth first traversals, each of which take $O(m)$ time. Non-uniform probabilities generation based on EDDBM can be done in $O(m)$ time because it uses one iteration of breadth first traversal. Thus, Algorithm 2 (*Betweenness - Ordering*(G, u, v)) takes $2 \cdot O(m) + 2 \cdot O(Tm) + 1 = O(Tm)$ time. Similarly, Algorithm 3 (k - *Betweenness - Ordering*(G, U)) takes $O(kTm + k \log k)$ time, where $O(kTm)$ factor is due to k times call to algorithm 1 and $O(k \log k)$ factor is for sorting.

In section 6.4.3, we note that only a constant number of iterations (T) are sufficient to provide efficient betweenness-ordering. Thus, we can write the time complexity of procedure *Betweenness - Ordering*(G, u, v) as $O(m)$. Similarly, if k is very smaller than n (number of total nodes) in the k -betweenness-ordering problem, we can assume k also as constant. In that case the time complexity of procedure k - *Betweenness - Ordering*(G, U) also becomes $O(m)$. Therefore, we name these algorithms as BOLT : betweenness ordering algorithms in almost linear time (in m) .

6 EXPERIMENTAL RESULTS

In this section, we discuss the experimental results achieved on extensive real world graphs and synthetic graphs. We have implemented the algorithms in C++. All the simulations were performed on a CentOS 6.5 machine with 2x (Xeon E5-2670V2(10 Core, 2.5Ghz)) processors and 96 GB RAM. The used system is a node in the Beowulf Cluster available at IIT Ropar with 30 similar nodes.

6.1 Data set**6.1.1 Real Networks**

We have picked a number of real world networks that are popular for betweenness computation and estimation in the related literature. We restricted the detailed analysis to the networks with nodes less than 40,000 due to computational constraints. We provide brief summary of these networks in Table 1 and [48], [49] can be referred for detailed description about the network data. For evaluating the performance of ordering by our algorithm, in addition to these networks, we have considered another 54 networks that cover most of the different networks available at [48], [50] of size (100 , 100k). We have considered collaboration networks, citation networks, communication network, social network, internet peer to peer network and many more. The columns of the Table 1 consist names of the network instances, size of networks or number of nodes (n), average degree of the nodes in the networks (Avg. Deg.), number of nodes with zero betweenness score (Z-BC) and type of networks respectively.

6.1.2 Synthetic Networks

We considered following types of artificial graphs:

- 1) **Random Graphs (ER)** [41]. For generating random graphs, we have considered the most extensively used random graph generation $G(n, p)$ model given by Erdos Renyi. The model takes as input the number of nodes n and a probability p . Then for each possible pair of different nodes, it puts an edge with a probability of p and outputs the generated graph.

TABLE 1
Considered Real World Networks

Instance name	n	Avg. Deg.	Z-BC	Network Type
as20000102 [48]	6474	3.88384	3682	Autonomous systems graph
Wiki-Vote [48]	7115	28.3238	2517	Social Network
wb-cs-stanford [49]	9435	5.81388	2814	Web Graph
CA-HepTh [48]	9877	5.25929	5291	Collaboration network
oregon1_010331 [48]	10670	4.12409	6285	Autonomous systems graph
PGPgiantcompo [49]	10680	4.55356	5663	Social Network
oregon1_010526 [48]	11174	4.18991	6520	Autonomous systems graph
CA-HepPh [48]	12008	19.735	6304	Collaboration network
CA-AstroPh [48]	18772	21.1006	8446	Collaboration network
p2p-Gnutella25 [48]	22687	4.82259	9348	Internet peer-to-peer network
as-22july06 [49]	22963	4.21861	11927	Internet Routers Network
CA-CondMat [48]	23133	8.07842	12635	Collaboration network
Cit-HepTh [48]	27770	25.3716	2345	Citation Network
Cit-HepPh [48]	34546	24.3662	2120	Citation Network
p2p-Gnutella30 [48]	36682	4.81588	16531	Internet peer-to-peer network
Email-Enron [48]	36692	10.0202	23710	Communication Network

We also referred this probability as edge existential probability in this paper.

- 2) **Scale-free Random Graphs (BA) [51].** For generating scale-free random graphs, we have considered the Albert Barabasi graph generation model. Throughout the paper, we denote it as $H(n, k)$. It takes as input the number of nodes (n) and an integer k . It starts with a complete graph of size k and keep adding random k different edges from new coming nodes to the existing nodes with probability as the normalized degree of old nodes. Thus this model is also referred many a time as preferential attachment model.

Table 2 summarizes the details about the picked artificial graphs where ER_n_x means $G(n, \frac{1}{n})$ and BA_n_x means $H(n, \lfloor \frac{n}{2} \rfloor)$. The columns of the Table 2 consist label to the networks, size of networks (n), other parameters (p/k) that needs to be fixed in the generation of synthetic networks, average degree of the nodes in the networks (Avg. Deg.), edges in the network and average number of nodes with zero betweenness score (Avg. Z-BC) respectively.

TABLE 2
Considered Synthetic Networks

Instance name	n	p/k	Avg. Deg.	Edge	Avg. Z-BC
ER_1k_2	1000	0.03162278	31.6976	15848	0
ER_1k_3	1000	0.01	10.01	5005	0.2
ER_1k_4	1000	0.00562341	5.616	2808	20.4
ER_1k_8	1000	0.00237137	2.4044	1202	306.4
ER_10k_2	10000	0.01	99.96312	499816	0
ER_10k_3	10000	0.002154435	21.48508	107425	0
ER_10k_4	10000	0.001	10.00588	50029	5.2
ER_10k_8	10000	0.000316228	3.1786	15893	1728.4
BA_1k_2	1000	16	31.488	15744	0
BA_1k_3	1000	5	9.95	4975	0
BA_1k_4	1000	3	5.982	2991	0
BA_1k_8	1000	2	3.992	1996	21.2
BA_10k_2	10000	50	99.5	497500	0
BA_10k_3	10000	11	21.9758	109879	0
BA_10k_4	10000	5	9.995	49975	0
BA_10k_8	10000	2	3.9992	19996	32.8

6.2 Performance Measurement Tools

In this section we will define various measures used to evaluate the performance of our model. These tools can be used for any betweenness computation or comparison algorithm to measure its performance.

6.2.1 For Betweenness Computation : Error and Average Error

Let a graph $G = (V, E)$ with $|V| = n$ is given. Let $BC^e(v)$ be the exact betweenness score of node v in the given graph. Let $BC^a(v)$ be the betweenness score of the same node v computed by Algorithm 1 using probabilities generated by EDDBM. Then, we define error in computation of betweenness score on node v same as Chehreghani [24]:

$$Er(v) = \frac{|BC^e(v) - BC^a(v)|}{BC^e(v)} \times 100$$

We define *average error* E in the computation of betweenness score of a set of nodes U , $U \subseteq V$, over a graph G as

$$E = \frac{\sum_{i \in U} Er(i)}{|U|}$$

where $|U|$ denotes the number of nodes in set U . To compute average error in the computation of betweenness score in a graph, we considered $U = \{v : v \in V \text{ and } BC^e(v) > 0\}$ throughout the paper. To find the average error in the betweenness computation for a node, we take mean of the error over five iterations. For synthetic graphs, we take mean of the average error over five such artificial graphs. *Number of iterations* used for computation of betweenness score is referred as the number of sampled nodes. We denote it by T .

6.2.2 For Betweenness Ordering : Efficiency and Relaxed Efficiency

Let n be the number of nodes in the considered graph. Then, $\binom{n}{2}$ different pairs of nodes are possible. Let $b_{ij} = 1$ if the result of betweenness comparison between node i and node j by our algorithm is correct, otherwise $b_{ij} = 0$. The efficiency of algorithm for betweenness-ordering of two nodes can be given as

$$\xi = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n b_{ij}}{\binom{n}{2}}.$$

In real world scenarios when two nodes possess very close betweenness ranks, error in the betweenness-ordering of those two nodes does not matter much. For example importance of the top central node or second top central node is very close. Thus, a relaxed version of the efficiency measure can be modelled. We take a threshold t and relax the ordering of nodes if the difference between the betweenness ranks of both the nodes is less or equal to t . By relaxing, we mean that we do not consider those pairs for measuring the efficiency of algorithm. Let P_t be the set of all pairs of nodes with betweenness rank difference greater than t and $|P_t|$ denotes the cardinality of set P_t . Let b_{ij} be a flag variable which gets value 1 for correct comparison and 0 otherwise. Then we can redefine the relaxed efficiency as:

$$\xi^t = \frac{\sum_{(i,j) \in P_t} b_{ij}}{|P_t|}.$$

At $t = 0$, $\xi^t = \xi$. Next, we name the algorithms picked for betweenness estimation and ordering analysis.

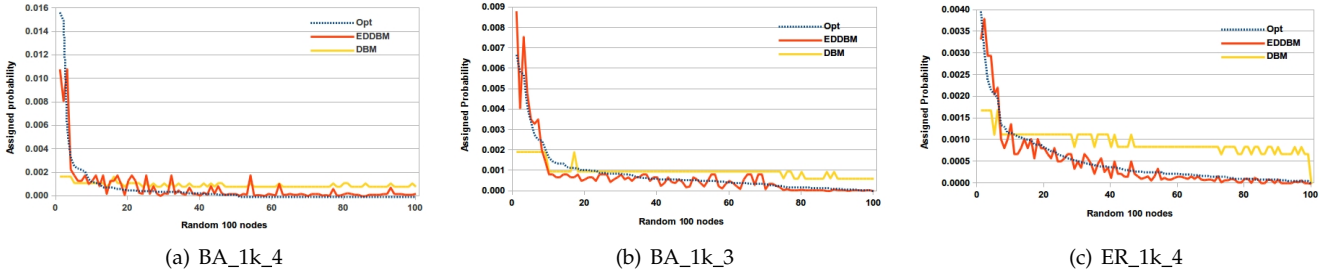


Fig. 4. Comparison of probabilities assigned by DBM and EDDBM vs the optimal model (opt) in four different artificial networks.

6.3 Considered Competitive Algorithms

In this section we mention the algorithms picked for comparative analysis with our approaches for betweenness ordering and betweenness estimation. We consider following labels *BP/B*, *LS*, *MC*, *Our*, *2-BC* for the Brandes and Pich's (time bounded version of Bader et al.'s [37]) uniform sampling based approximation algorithm, Geisberger et al.'s [38] linear scaling based algorithm, Chehreghani [24]'s recent algorithm, our algorithm and Gkorou et al.'s [39] k -betweenness algorithm with $k = 2$ respectively. The first four algorithms are node sampling based (probabilistic) algorithm and we fix equal number of samples for each one of these algorithms. The last, 2-BC algorithm is a deterministic algorithm and takes a lot more time than the sampling based algorithms. We have considered this algorithm to show that even with very few samples (very less time), our algorithm outperforms this huge time taking deterministic algorithm most of the times. We also considered the recent path sampling based Riondato and Kornaropoulos's [40] randomized algorithm but due to its bad performance, we skip the results. Their algorithm is theoretically sound but in the small time fixed by us, it does not perform well for estimation or ordering on any considered network. Next, we see various plots to analyze our model and algorithms.

6.4 Plots

In this section we evaluate the performance of EDDBM for estimation and ordering through various plots. Next, with the help of different plots on synthetic and real world networks, we compare the accuracy of EDDBM in comparison to DBM.

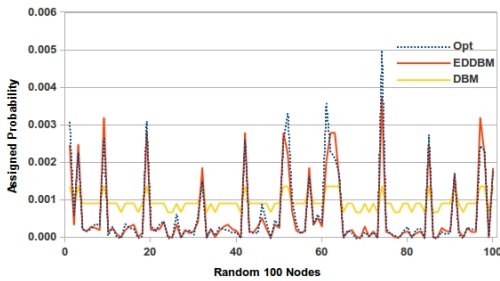


Fig. 5. Comparison of probabilities assigned by DBM and EDDBM vs the optimal model (Opt) in ER_1k_3.

6.4.1 Comparison of probabilities assigned by DBM, EDDBM and optimal model

Plots in this section analyze experimentally the approximation performance of EDDBM and shows that EDDBM generates probabilities very close to the optimal probabilities. These plots also compare EDDBM with DBM. All the plots in Fig. 4 and Fig. 5 are drawn for 4 different synthetic networks which are picked from Table 2.

For generating the plots, we picked a random node from each network. With the assumption to estimate betweenness of this node, we assign probabilities to all the nodes in the network using EDDBM model, optimal sampling model (Opt), and DBM. Each network consists a large number of nodes. To draw a clear plot, we randomly picked 100 nodes and plotted probabilities assigned by the above mentioned sampling models for only these 100 randomly picked nodes. The x-axis represents the 100 nodes that are picked and the y-axis represent the probabilities assigned by DBM, EDDBM and the optimal model (Opt). In the first three plots in Fig. 4, we sorted the randomly picked 100 nodes in descending order based on the optimal probabilities assigned to them before plotting. The next one plot in Fig. 5 is without such sorting process.

In both the figures Fig. 4 and Fig.5, it is easy to observe that EDDBM is much better than DBM and EDDBM generates probabilities very close to the optimal probabilities. We also note in Fig.5 that the plot of probabilities by EDDBM achieves similar characteristic peaks as the plot of optimal probabilities get. This infers that, unlike DBM, EDDBM identifies the nodes with high contribution and assigns large probabilities to them. EDDBM also focuses on the nodes contributing very less and tries to assign smaller probabilities to them which was not well handled by DBM.

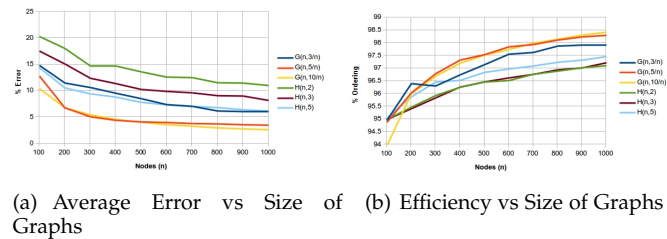


Fig. 6. Average Error and Efficiency vs Size of Synthetic (ER and BA) Graphs

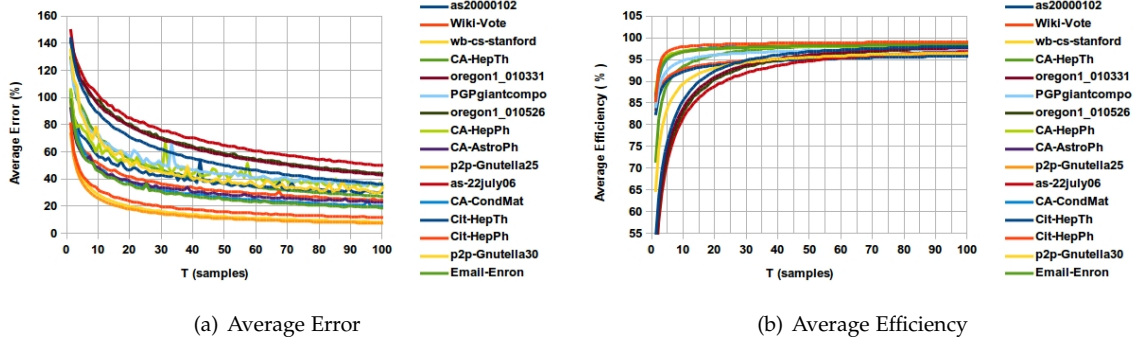


Fig. 9. Iterative performance of our approach on considered real world networks.

6.4.2 Average Error and Efficiency vs Size of Graphs (n)

In this section we plot the average error in the computation of betweenness score and efficiency in ordering the nodes based on betweenness scores in a graph with respect to the size of graph (number of nodes in that graph). We generated graphs with $n = 100$ to $n = 1000$ with a step of 100. For each n , we generate 5 graphs and averaged the average error and efficiency over all 5 graphs. Fig. 6 contains the plots. Plot in Fig. 6(a) depicts the change in average error and plot in Fig. 6(b) depicts the change in efficiency while changing the size of graphs but keeping average degree as a constant. We plotted the results for ER graphs with average degree = $\{3, 5, 10\}$ and for BA graphs with average degree = $\{4, 6, 10\}$. From the plots, we can infer that the average error in computation of betweenness score decreases and efficiency in ordering the nodes based on betweenness score increases with increase in the size of graph.

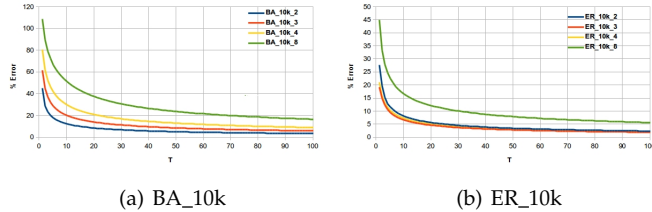


Fig. 7. Iterative average error performance of our approach on considered synthetic graphs.

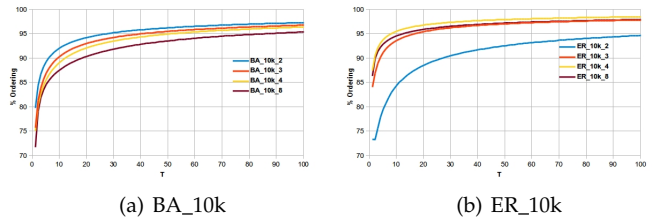


Fig. 8. Iterative average ordering performance of our approach on considered synthetic graphs.

6.4.3 Average Error and Efficiency vs Number of Sampled Nodes (T)

In this section we plot the average error in the computation of betweenness centrality using EDDBM and efficiency

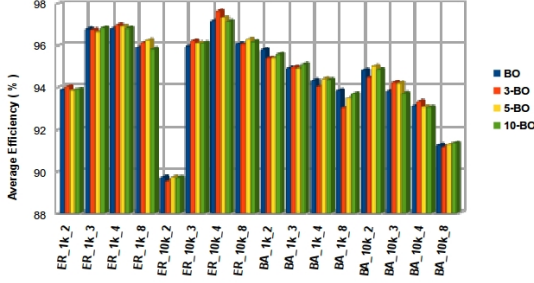
in ordering the nodes based on betweenness score using Algorithm 2, when the number of sampled nodes (no of iterations) were $T = X$. These plots were drawn to inspect what value of T that can suffice for good result i.e. betweenness computation with less error and ordering with high efficiency. In each of the Fig. 7 and Fig. 8, there are two plots. One for the ER graphs and the other for BA graphs of size 10000 that are mentioned in Table 2. Fig. 7 is the plots of change in the average error vs T and Fig. 8 is the plots of change in the efficiency vs T on considered synthetic graphs. Fig. 9 is the plots of change in the average error vs T and change in the efficiency vs T on considered real world graphs.

For simplicity, we start all the plots in this section from $X = 1$. In the plots in this section, the average error reduces and average efficiency increases very sharply when X varies from 1 to 15 or 25. After $X = 25$ there is very small change in the average error and efficiency. Due to our focus on a linear time betweenness-ordering algorithm, we focus on the iterative efficiency(ordering) performance plot on real world networks given in Fig. 9. It is notable that at $T = 25$, on all considered real world graphs, efficiency reaches beyond 90% and for most of them even beyond 95%. Thus, we set $T = 25$ to achieve experimental results in this paper. By reducing T to a constant, one can suspect that the error might increase in larger graphs but in larger graphs our model performs much better. The reason is given in previous section that the average error decreases and efficiency increases with the increase in number of nodes which neutralizes the effect of increase in error by keeping T as constant.

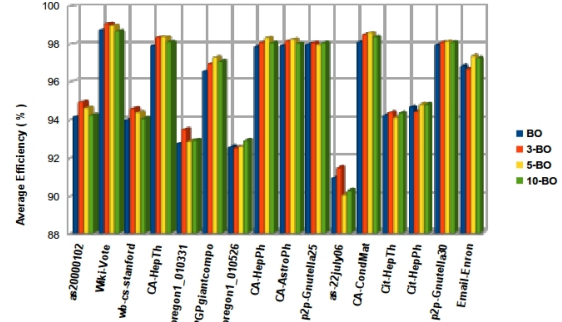
6.4.4 Average efficiency in k -betweenness-ordering vs k

In this section, we plot the the average efficiency of our approach for ordering k nodes with respect to considered synthetic and real world graphs. The average efficiency for k -betweenness ordering is calculated as following. In a given graph, we randomly pick a set of k nodes and calculate the efficiency in ordering these k nodes based on the formula given in section 6.2.2. We do this for 1000 iterations and take average of the efficiency in these 1000 iterations, called average efficiency.

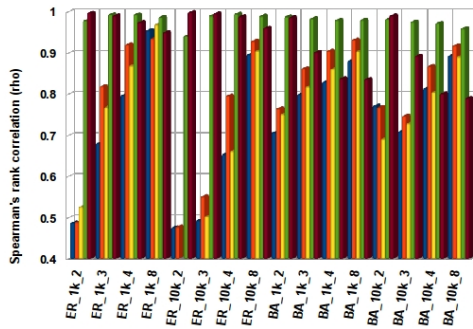
The plots in Fig. 10 show the average efficiency of Algorithm 3 for k -betweenness-ordering on various artificial networks (Fig. 10(a)) and real world networks (Fig. 10(b)). Label



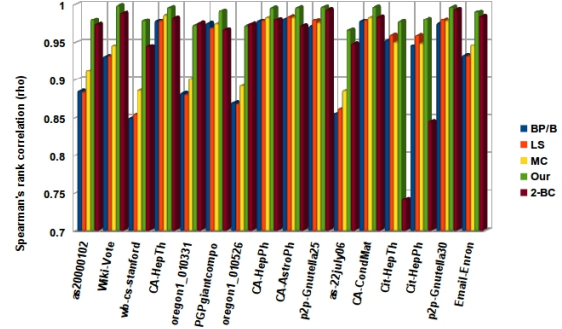
(a) On artificial networks



(b) On real world networks

Fig. 10. Average k -betweenness-ordering efficiency for different values of k 

(a) On artificial networks



(b) On real world networks

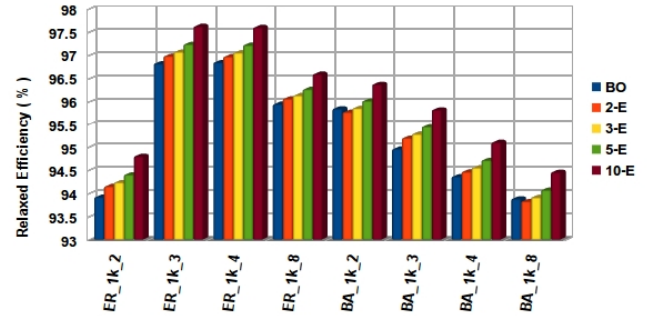
Fig. 11. Average spearman rank correlation (ρ) of various algorithms in considered artificial and real world networks.

BO, denotes the efficiency of betweenness-ordering. Labels 3-BO, 5-BO, and 10-BO denotes the average efficiency of our approach for k -betweenness-ordering on $k = 3, 5, 10$ respectively. Plots show that the performance of our approach (Algorithm 3) is nearly same for k -betweenness-ordering as the performance of Algorithm 2 for the betweenness-ordering (ordering of two nodes).

6.4.5 Correlation in ordering / Average efficiency vs Size of Graphs (n)

In this section we plot the average *Spearman rank correlation* (ρ) between the results produced by algorithm 2 and other node sampling based approaches on artificial and real world graphs. ρ is a standard ranking correlation that measures the similarity between ordering of two ranking algorithms. The plots are given in Fig. 11.

Correlation of our approach is very high (very close to 1) for almost all networks. Our approach outperforms all node sampling based algorithms. In very dense networks, 2-BC sometimes produces better result than our approach but it should be noted that the difference in correlations are minute even when 2-BC takes a lot more time than our approach. In sparse networks, our approach produces much better results than 2-BC in very small amount of time.

Fig. 12. Average relaxed efficiency for different values of t in some artificial networks

6.4.6 Average Relaxed efficiency (ξ^t) vs t

In this section we will inspect the plot between the relaxed efficiency ξ^t and t on some artificial networks of 1k nodes. Similar results are achieved on all considered networks but due to page limit, we skip the plots for other networks. The relaxed efficiency is computed by the formula given in section 6.2.2. We vary t for $t=2,3,5,10$. The plots are compiled in Fig. 12. For each t , we generate 5 artificial networks and then calculate the relaxed efficiency on each graph and average them to get average relaxed

efficiency. We plotted average relaxed efficiency for both type of considered artificial graphs (ER and BA). BO, 2-E, 3-E, 5-E, 10-E are the labels assumed for average relaxed efficiency at $t=0,2,3,5,10$ respectively. The plots demonstrate that the efficiency increases with an increase in t i.e. in real world situation where relaxation is allowed in ordering, our algorithm will perform much better than its usual performance. The relaxation in ordering means that betweenness ordering of the nodes with approximately same betweenness rank is ignored and only the betweenness ordering of nodes with difference greater than the threshold value (t) in their betweenness ranks are considered. Thus, we can say that our algorithm results ordering very close to the exact betweenness-ordering.

In next section, we will discuss the betweenness estimation and ordering results achieved for considered synthetic networks and a number of real world networks.

6.5 Average error and Efficiency in Graphs

Here, we discuss and compare the results achieved by our (BOLT) and other algorithms that are mentioned in section 6.3 on considered synthetic networks and several real world networks. The tables are provided in the appendix as supplementary material.

6.5.1 Average error and Efficiency in synthetic graphs

In this section, we analyze the results over synthetic graphs mentioned in section 6.1.2. The average of average error is 88.138, 58.063, 67.719, **12.564**, and 86.548 percentage by BP/B, LS, MC, **Our**, and 2-BC respectively. The average of average efficiency is 78.913, 81.666, 79.455, **94.684**, and 87.383 percentage by BP/B, LS, MC, **Our**, and 2-BC respectively. It is easy to observe that our algorithm outperforms all mentioned sampling based algorithms for both, estimation and ordering by a huge margin. 2-BC algorithm performs better for ordering but not for estimation in very dense graph. It is because of very small average distance between nodes. But it should be noted that 2-BC takes several folds more time than our algorithm. In moderately dense or sparse, our algorithm even outperforms 2-BC with a big margin. Our model performs relatively better in dense graphs than in sparse graphs.

6.5.2 Average error and Efficiency in real-world graphs

This section presents and discusses the simulation results on various real networks considered in section 6.1.1. After extracting the networks, we converted the networks into unweighted undirected networks, if required. Then we removed multi-edges, self-loops and isolated nodes if existing. The average of average error is 109.469, 69.210, 98.603, **47.063**, and 95.289 percentage by BP/B, LS, MC, **Our**, and 2-BC respectively. The average of average efficiency is 84.803, 84.841, 87.294, **95.861**, 92.317 percentage by BP/B, LS, MC, **Our**, and 2-BC respectively. Our algorithm again superseded all other considered algorithms for betweenness estimation. The estimation results can be improved by increasing the value of T that is considered 25 for all computations.

The betweenness-ordering results for $T = 25$ and $T = 50$ are calculated on 16 networks mentioned in section 6.1.1 and

another 54 picked networks. We achieve average efficiency 96.586 and 97.632 percentage with standard deviation of 2.384 and 1.447 percentage for $T = 25, 50$ respectively ignoring three special networks. The details of the other 54 networks and ordering results on all real-world networks are provided in the appendix as supplementary material.

7 CONCLUSION AND FURTHER WORK

In this paper, we coin a new problem called the betweenness-ordering-problem and address its importance with real world examples and provide a feasible and practical almost linear time solution to it. By our problem statement, betweenness-ordering problem refers to the ordering of two nodes. We also solve the general version of the betweenness-ordering problem for k nodes. We present the proof of concept of our technique by applying it to real world graphs as well as synthetic ones. To the best of our knowledge, this is the first of its kind study addressing the "ordering problem" in centrality measures. While our work is a first attempt to provide a solution to the centrality-ordering-problem for the betweenness measure, we feel this should lead to the asking and answering of this question across several popular measures that have seen its applications in diverse areas.

- Our model performs very well on both real and synthetic networks. The formulation of EDDBM is based on the analysis of random graphs. Random graphs does not possess high clustering coefficient and thus this model does not perform well on the graphs with high clustering coefficient. In highly clustered graphs, a better model is desirable. An interesting problem would be to tune our algorithm so that the clustering has no effect on the results.
- One can attempt to ask a similar question for the Pagerank-ordering-problem, closeness-ordering-problem or any centrality-ordering-problem. Any attempt to address these problems in the same spirit as our addressing the betweenness-ordering-problem would collectively be a great contribution to applied sciences where centrality measures are being increasingly applied.

ACKNOWLEDGMENT

The authors would like to thank the IIT Ropar HPC committee for providing the resources to perform experiments.

REFERENCES

- [1] U. Brandes and T. Erlebach, *Network analysis: methodological foundations*. Springer, 2005, vol. 3418.
- [2] M. Newman, *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc., 2010.
- [3] M. O. Jackson, *Social and Economic Networks*. Princeton, NJ, USA: Princeton University Press, 2008.
- [4] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 990–998.
- [5] P. Hage and F. Harary, "Eccentricity and centrality in networks," *Social Networks*, vol. 17, no. 1, pp. 57 – 63, 1995.
- [6] E. Welzl, *Smallest enclosing disks (balls and ellipsoids)*. Springer, 1991.

- [7] S. Kintali, "Betweenness centrality: Algorithms and lower bounds," *arXiv preprint arXiv:0809.1906*, 2008.
- [8] M. Agarwal, R. R. Singh, S. Chaudhary, and S. Iyengar, "An efficient estimation of a nodes betweenness," in *Complex Networks VI*. Springer, 2015, pp. 111–121.
- [9] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.
- [10] J. M. Anthonisse, "The rush in a directed graph," *Stichting Mathematisch Centrum. Mathematische Besliskunde*, no. BN 9/71, pp. 1–10, 1971.
- [11] S. Narayanan, "The betweenness centrality of biological networks," Ph.D. dissertation, Virginia Polytechnic Institute and State University, 2005.
- [12] M. P. Joy, A. Brock, D. E. Ingber, and S. Huang, "High-betweenness proteins in the yeast protein interaction network," *BioMed Research International*, vol. 2005, no. 2, pp. 96–103, 2005.
- [13] A. Tizghadam and A. Leon-Garcia, "Betweenness centrality and resistance distance in communication networks," *Network, IEEE*, vol. 24, no. 6, pp. 10–16, 2010.
- [14] T.-C. Lu, Y. Zhang, D. L. Allen, and M. A. Salman, "Design for fault analysis using multi-partite, multi-attribute betweenness centrality measures," 2011.
- [15] S. P. Borgatti and X. Li, "On social network analysis in a supply chain context," *Journal of Supply Chain Management*, vol. 45, no. 2, pp. 5–22, 2009.
- [16] K. J. Mizgier, M. P. Jüttner, and S. M. Wagner, "Bottleneck identification in supply chain networks," *International Journal of Production Research*, vol. 51, no. 5, pp. 1477–1490, 2013.
- [17] S. Derrible, "Network centrality of metro systems," *PloS one*, vol. 7, no. 7, p. e40575, 2012.
- [18] R. Carvalho, L. Buzna, F. Bono, E. Gutiérrez, W. Just, and D. Arrowsmith, "Robustness of trans-european gas networks," *Physical review E*, vol. 80, no. 1, p. 016106, 2009.
- [19] V. Rosato, L. Issacharoff, F. Tiriticco, S. Meloni, S. Porcellinis, and R. Setola, "Modelling interdependent infrastructures using interacting dynamical models," *International Journal of Critical Infrastructures*, vol. 4, no. 1, pp. 63–79, 2008.
- [20] R. Kinney, P. Crucitti, R. Albert, and V. Latora, "Modeling cascading failures in the north american power grid," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 46, no. 1, pp. 101–107, 2005.
- [21] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, "Catastrophic cascade of failures in interdependent networks," *Nature*, vol. 464, no. 7291, pp. 1025–1028, 2010.
- [22] G. Lin, Z. Di, and Y. Fan, "Cascading failures in complex networks with community structure," *International Journal of Modern Physics C*, vol. 25, no. 05, 2014.
- [23] A. E. Motter and Y.-C. Lai, "Cascade-based attacks on complex networks," *Physical Review E*, vol. 66, no. 6, p. 065102, 2002.
- [24] M. H. Chehreghani, "An efficient algorithm for approximate betweenness centrality computation," *The Computer Journal*, p. bxu003, 2014.
- [25] R. W. Floyd, "Algorithm 97: shortest path," *Communications of the ACM*, vol. 5, no. 6, p. 345, 1962.
- [26] S. Warshall, "A theorem on boolean matrices," *Journal of the ACM (JACM)*, vol. 9, no. 1, pp. 11–12, 1962.
- [27] U. Brandes, "A faster algorithm for betweenness centrality," *The Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, 2001.
- [28] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.
- [29] A. E. Sariyüce, E. Saule, K. Kaya, and Ü. V. Çatalyürek, "Shattering and compressing networks for betweenness centrality," in *SIAM Data Mining Conference (SDM)*. SIAM, 2013.
- [30] M.-J. Lee, J. Lee, J. Y. Park, R. H. Choi, and C.-W. Chung, "Qube: A quick algorithm for updating betweenness centrality," in *Proceedings of the 21st International Conference on World Wide Web*, ser. WWW '12. New York, NY, USA: ACM, 2012, pp. 351–360.
- [31] O. Green, R. McColl, and D. Bader, "A fast algorithm for streaming betweenness centrality," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*, Sept 2012, pp. 11–20.
- [32] M. Kas, M. Wachs, K. M. Carley, and L. R. Carley, "Incremental algorithm for updating betweenness centrality in dynamically growing networks," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '13. New York, NY, USA: ACM, 2013, pp. 33–40.
- [33] K. Goel, R. R. Singh, S. Iyengar *et al.*, "A faster algorithm to update betweenness centrality after node alteration," in *Algorithms and Models for the Web Graph*. Springer, 2013, pp. 170–184.
- [34] M. Nasre, M. Pontecorvi, and V. Ramachandran, "Betweenness centrality-incremental and faster," *arXiv preprint arXiv:1311.2147*, 2013.
- [35] D. Eppstein and J. Wang, "Fast approximation of centrality," *J. Graph Algorithms Appl.*, vol. 8, pp. 39–45, 2004.
- [36] U. Brandes and C. Pich, "Centrality estimation in large networks," *International Journal of Bifurcation and Chaos*, vol. 17, no. 07, pp. 2303–2318, 2007.
- [37] D. A. Bader, S. Kintali, K. Madduri, and M. Mihail, "Approximating betweenness centrality," in *Proceedings of the 5th International Conference on Algorithms and Models for the Web-graph*, ser. WAW'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 124–137.
- [38] R. Geisberger, P. Sanders, and D. Schultes, *Better Approximation of Betweenness Centrality*, 2008, ch. 8, pp. 90–100.
- [39] D. Gkorou, J. Pouwelse, D. Epema, T. Kielmann, M. van Kreveld, and W. Niessen, "Efficient approximate computation of betweenness centrality," in *16th annual conf. of the Advanced School for Computing and Imaging (ASCI 2010)*, 2010.
- [40] M. Riondato and E. M. Kornaropoulos, "Fast approximation of betweenness centrality through sampling," in *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 2014, pp. 413–422.
- [41] P. Erdos and A. Renyi, "On random graphs i," *Publ. Math. Debrecen*, vol. 6, pp. 290–297, 1959.
- [42] X. Wang, "Deciding on the type of the degree distribution of a graph (network) from traceroute-like measurements," 2011.
- [43] R. Van Der Hofstad, "Random graphs and complex networks," Available on <http://www.win.tue.nl/rhofstad/NotesRGCN.pdf>, 2009.
- [44] C. Bauckhage, K. Kersting, and B. Rastegarpanah, "The weibull as a model of shortest path distributions in random networks," in *Proc. Int. Workshop on Mining and Learning with Graphs*, Chicago, IL, USA, 2013.
- [45] T. W. Valente, K. Coronges, C. Lakon, and E. Costenbader, "How correlated are network centrality measures?" *Connections (Toronto, Ont.)*, vol. 28, no. 1, p. 16, 2008.
- [46] C.-Y. Lee, "Correlations among centrality measures in complex networks," *arXiv preprint physics/0605220*, 2006.
- [47] D. E. Knuth, *The art of computer programming: sorting and searching*. Pearson Education, 1998, vol. 3.
- [48] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," <http://snap.stanford.edu/data>, Jun. 2014.
- [49] T. A. Davis and Y. Hu, "The university of florida sparse matrix collection," *ACM Transactions on Mathematical Software (TOMS)*, vol. 38, no. 1, p. 1, 2011.
- [50] V. Batagelj and A. Mrvar, "Pajek datasets," *Web page* <http://vlado.fmf.uni-lj.si/pub/networks/data>, 2006.
- [51] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.



Manas Agarwal: Manas Agarwal is an undergraduate student, currently pursuing Integrated Master of Science in Applied Mathematics from Indian Institute of Technology Roorkee, India. He currently holds Departmental Rank 1 and is a KVPY (Kishore Vaigyanik Protsahan Yojana) scholar. He is a fellow of the Indian Academy of Sciences, Bangalore. His current interests are in Number Theory, Graph Theory and various other areas of Discrete Mathematics and their applications in design and analysis of algorithms.



Rishi Ranjan Singh: Mr. Singh is a PhD Scholar in the Department of Computer Science and Engineering at IIT Ropar. He is interested in Approximation Algorithms for Aehicle Routing problems, Social and Complex Network, Network Analysis and Operations Research domain.



Shubham Chaudhary: Shubham Chaudhary is currently a third year student of Integrated Master of Science in Applied Mathematics at Indian Institute of Technology Roorkee, India. He is a KVPY (Kishore Vaigyanik Protsahan Yojana) scholar and was awarded research fellowship by Indian Academy of Sciences, Bangalore. His areas of interest are Graph Theory, Number Theory and Combinatorics.



S.R.S. Iyengar: Mr. Iyengar is an assistant professor in the Department of Computer Science and Engineering at IIT Ropar. He received PhD degree from Indian Institute of Science, Bangalore. His research interests include social networks, crowdsourced knowledge building and computational social sciences.

APPENDIX A

A.0.1 Average error and Efficiency in synthetic graphs

In this section, we present results over the considered synthetic graphs. The results are summarized in Table 3 and Table 4. The first column in both of the tables consist label of the network instance. The next 5 columns in Table 3 are the estimation performance, average error in estimating betweenness, by various algorithms. In Table 4, last 5 columns are the ordering performance results, average efficiency in ordering 2 nodes based on betweenness score, by various algorithms.

TABLE 3
Average error in artificial networks when $T = 25$

Instance	BP/B	LS	MC	Our	2-BC
ER_1k_2	66.291	50.603	43.433	5.425	53.982
ER_1k_3	70.015	37.699	44.585	5.345	95.880
ER_1k_4	67.505	42.140	44.097	7.505	99.033
ER_1k_5	63.252	62.835	37.646	18.257	99.411
ER_10k_2	113.418	90.286	82.129	5.281	53.826
ER_10k_3	89.656	52.754	75.481	4.381	98.045
ER_10k_4	92.911	47.576	72.433	4.793	99.699
ER_10k_5	83.714	65.528	66.032	14.084	99.933
BA_1k_2	83.915	64.922	56.750	7.987	52.232
BA_1k_3	82.078	47.664	61.647	12.487	93.155
BA_1k_4	84.845	45.439	63.220	17.155	97.567
BA_1k_8	82.732	48.137	66.511	23.427	98.868
BA_10k_2	123.117	100.389	99.124	8.208	48.404
BA_10k_3	100.749	65.610	90.002	13.017	95.667
BA_10k_4	103.962	54.955	89.731	18.844	99.179
BA_10k_8	102.052	52.474	90.697	34.829	99.892

TABLE 4
Average Efficiency in artificial networks when $T = 25$

Instance	BP/B	LS	MC	Our	2-BC
ER_1k_2	66.428	67.685	68.193	93.927	98.346
ER_1k_3	74.186	80.938	78.054	96.828	95.481
ER_1k_4	80.398	88.441	83.943	96.854	90.025
ER_1k_5	90.980	89.678	92.646	95.954	83.221
ER_10k_2	67.173	67.310	63.935	89.760	99.295
ER_10k_3	67.504	69.518	67.817	96.023	97.256
ER_10k_4	73.776	79.966	74.018	97.215	92.901
ER_10k_5	87.716	88.738	87.845	96.128	82.815
BA_1k_2	77.803	79.644	79.024	95.855	95.583
BA_1k_3	80.854	84.555	81.845	94.965	86.927
BA_1k_4	82.889	87.979	84.245	94.372	79.543
BA_1k_8	86.115	89.199	86.731	93.889	70.505
BA_10k_2	80.993	81.583	76.898	94.873	96.455
BA_10k_3	77.458	78.496	77.827	93.859	86.472
BA_10k_4	81.785	84.967	81.571	93.161	79.137
BA_10k_8	86.549	87.962	86.687	91.289	64.161

TABLE 5
Average error in real world networks when $T = 25$

Instance	BP/B	LS	MC	Our	2-BC
as20000102	110.466	70.932	100.000	66.265	98.683
Wiki-Vote	117.143	78.948	104.930	22.355	93.650
wb-cs-stanford	135.742	126.957	123.029	52.811	62.970
CA-HepTh	98.828	57.641	77.841	35.233	96.075
oregon1_010331	134.948	76.411	109.810	74.394	98.711
PGPgiantcompo	96.975	52.619	93.898	54.883	96.212
oregon1_010526	115.421	77.324	115.177	74.300	98.722
CA-HepPh	95.320	55.769	86.426	50.714	96.254
CA-AstroPh	98.069	54.855	91.558	36.938	97.947
p2p-Gnutella25	102.861	63.462	83.838	16.760	99.962
as-22july06	131.187	82.954	115.373	80.428	98.965
CA-CondMat	95.170	53.324	89.418	34.532	96.333
Cit-HepTh	104.839	59.056	96.572	44.896	98.253
Cit-HepPh	109.118	58.523	95.126	39.311	98.949
p2p-Gnutella30	100.685	64.435	90.594	18.407	99.957
Email-Enron	104.747	74.159	104.052	50.781	92.997

TABLE 6
Average Efficiency in real world networks when $T = 25$

Instance	BP/B	LS	MC	Our	2-BC
as20000102	76.398	76.336	80.574	94.224	93.390
Wiki-Vote	81.246	81.296	84.614	98.779	96.037
wb-cs-stanford	63.901	63.685	71.806	94.041	91.597
CA-HepTh	93.160	92.910	95.442	97.951	94.732
oregon1_010331	75.061	74.862	78.706	92.793	93.659
PGPgiantcompo	92.374	90.398	92.337	96.606	92.401
oregon1_010526	72.859	72.453	77.153	92.594	93.461
CA-HepPh	93.092	93.064	94.650	97.886	94.629
CA-AstroPh	93.778	94.178	94.496	97.934	93.824
p2p-Gnutella25	92.922	93.918	93.741	98.019	94.969
as-22july06	70.131	70.965	75.438	90.969	90.567
CA-CondMat	93.234	93.131	94.671	98.147	95.144
Cit-HepTh	91.014	91.204	91.043	94.270	78.306
Cit-HepPh	90.171	91.209	90.744	94.695	83.741
p2p-Gnutella30	93.428	93.980	94.120	97.989	95.347
Email-Enron	84.086	83.866	87.182	96.876	95.275

A.0.2 Average error and Efficiency in real-world graphs

This section presents and discusses the simulation results on various real-world networks. After extracting the networks, we converted the networks into unweighted undirected networks, if required. Then we removed multi-edges, self-loops and isolated nodes if existing. The results obtained are summarized in the Table 5 and Table 6. The columns are in similar order as in the Table 3 and Table 4 respectively. Here, Table 5 summarizes estimation performance results while Table 6 contains ordering performance results.

Further we presents and discusses the ordering results by our algorithm on more real-world networks. The results obtained are summarized in the Table 7 and Table 8. The first five columns of the Table 7 contain serial number, name of the network instances, size of networks (n), average degree of the nodes in the networks (Avg. Deg.), number of nodes with zero betweenness score (Z-BC) respectively. Next column contains the average efficiency of our algorithm for ordering when $T = 25$ and all $\binom{n}{2}$ pairs are considered for calculating the efficiency. Next column contains the average efficiency of our algorithm for ordering when $T = 25$ and only the pairs that consist at least one node with nonzero betweenness scores, are considered for computing the efficiency. The next two columns are same as the columns 6-7 except the efficiency is computed when $T=50$ (50 samples) is set in our algorithm. The picked networks almost cover the different networks data-sets available at [48], [50] of size (100 , 100k). The results show that the efficiency of our algorithm is very high and very close to the exact ordering for only a constant number of samples.

On three of the real-world networks that are picked, the ordering result were not good. The results are mentioned in table 8. The reason of this bad performance is a special properties of these networks. In these networks, most of the nodes share same betweenness score. Our method probabilistically estimates the betweenness score. The defined efficiency formula considers when two nodes share same actual betweenness score, they get same rank and the efficiency only increases if the considered estimation algorithm also assigns both of the nodes exactly same score. But by any probabilistic algorithm, even if it is very efficient, due to the probabilistic nature, it is very less probable (nearly impossible) that it will be able to assign exactly same score. And thus our algorithms gets lower efficiency though the average errors in the estimation were very small.

TABLE 7
Average Efficiency (in %) of our algorithm on real world networks setting $T = 25, 50$

S.N.	Instance name	n	Avg. Deg.	Z-BC	Ordering_25	Ordering_25 nz	Ordering_50	Ordering_50 nz
1	as20000102	6474	3.884	3682	96.092	94.224	97.949	96.968
2	Wiki-Vote	7115	28.324	2517	98.932	98.779	99.174	99.056
3	wb-cs-stanford	9435	5.814	2814	94.571	94.041	96.186	95.814
4	CA-HepTh	9877	5.259	5291	98.539	97.951	98.856	98.396
5	oregon1_010331	10670	4.124	6285	95.293	92.793	97.643	96.390
6	PGPgiantcompo	10680	4.554	5663	97.560	96.606	98.011	97.233
7	oregon1_010526	11174	4.190	6520	95.115	92.594	97.407	96.068
8	CA-HepPh	12008	19.735	6304	98.468	97.886	98.672	98.166
9	CA-AstroPh	18772	21.101	8446	98.352	97.934	98.674	98.338
10	p2p-Gnutella25	22687	4.823	9348	98.355	98.019	98.809	98.566
11	as-22july06	22963	4.219	11927	93.405	90.969	96.326	94.969
12	CA-CondMat	23133	8.078	12635	98.700	98.147	99.000	98.575
13	Cit-HepTh	27770	25.372	2345	94.311	94.270	95.286	95.253
14	Cit-HepPh	34546	24.366	2120	94.715	94.695	95.647	95.631
15	p2p-Gnutella30	36682	4.816	16531	98.397	97.989	98.831	98.533
16	Email-Enron	36692	10.020	23710	98.181	96.876	98.998	98.280
17	as19990829	103	4.641	43	98.835	98.593	99.101	98.915
18	facebook_combined	4039	43.691	342	96.396	96.370	97.242	97.222
19	CA-GrQcNew	5242	5.526	3236	98.957	98.315	99.143	98.616
20	P2p-Gnutella04	10876	7.355	2484	97.634	97.504	98.280	98.186
21	oregon2_010331	10900	5.721	6096	96.706	95.207	98.311	97.543
22	Oregon2_010526	11461	5.712	6290	96.345	94.770	98.042	97.199
23	P2p-Gnutella24	26518	4.930	11014	98.214	97.842	98.706	98.436
24	P2p-Gnutella31	62586	4.726	28829	98.241	97.768	98.708	98.360
25	Soc-Epinions1	75879	10.694	41048	98.006	97.181	98.529	97.921
26	Slashdot0811	77360	12.130	30164	97.575	97.140	98.179	97.853
27	Slashdot0902	82168	12.273	30855	97.493	97.081	98.151	97.848
28	GD99_c	105	2.286	36	95.190	94.563	95.769	95.217
29	GD98_b	121	2.182	75	97.702	96.281	98.014	96.785
30	Journals	124	96.323	0	94.443	94.443	95.951	95.951
31	GD96_d	180	2.533	58	92.007	91.094	92.831	92.011
32	GD01_a	311	4.116	121	97.743	97.343	98.169	97.845
33	USAir97	332	12.807	135	98.901	98.685	99.189	99.029
34	GD00_a	352	2.182	177	98.588	98.112	98.835	98.442
35	SmallW	396	5.020	228	98.522	97.791	99.244	98.871
36	GD97_c	452	2.035	395	99.893	99.548	99.920	99.661
37	Erdos971	472	5.568	168	97.643	97.303	98.130	97.860
38	Erdos981	485	5.695	171	97.690	97.363	98.151	97.890
39	Erdos991	492	5.760	173	97.622	97.287	98.213	97.962
40	GD00_c	638	3.197	273	97.729	97.221	98.391	98.031
41	GD01_Acap	953	1.343	763	99.595	98.871	99.723	99.228
42	Roget	1022	7.139	86	95.911	95.882	96.894	96.872
43	SmaGri	1059	9.284	251	97.350	97.193	98.118	98.006
44	GD96_a	1096	3.060	4	91.707	91.707	93.822	93.822
45	GD06_Java	1538	10.165	394	95.931	95.645	97.439	97.259
46	Csphd	1882	1.849	1306	99.481	99.000	99.548	99.129
47	Yeast	2361	5.630	952	98.288	97.956	98.742	98.498
48	ODLIS	2909	11.260	567	96.305	96.160	97.321	97.215
49	SciMet	3084	6.744	894	97.687	97.475	98.306	98.151
50	Kohonen	4470	5.690	1671	96.158	95.534	97.365	96.937
51	EPA	4772	3.734	2716	98.679	98.047	99.125	98.706
52	UspowerGrid	4941	2.669	1447	95.819	95.427	96.576	96.255
53	Erdos972	5488	2.582	4321	99.713	99.246	99.796	99.464
54	Erdos982	5822	2.533	4647	99.718	99.222	99.812	99.483
55	Erdos992	6100	2.464	4911	99.713	99.184	99.822	99.495
56	Zewail	6752	16.049	827	96.440	96.385	97.203	97.161
57	Erdos02	6927	2.446	5640	95.243	85.887	98.308	94.981
58	Geom	7343	3.241	5677	99.379	98.456	99.678	99.200
59	EVA	8497	1.580	7656	99.847	99.188	99.873	99.323
60	Lederberg	8843	9.393	2417	96.111	95.797	97.148	96.918
61	California	9664	3.305	5958	98.866	98.170	99.270	98.823
62	FA	10617	12.016	4235	98.324	98.007	98.715	98.472
63	foldoc	13356	13.697	1	94.982	94.982	96.410	96.410
64	EAT_RS	23219	26.266	2860	97.376	97.336	98.072	98.043
65	EAT_SR	23219	26.266	2861	97.364	97.324	98.071	98.041
66	Dictionary28	52652	3.382	27847	98.498	97.915	98.778	98.303
67	Wordnet3	82670	2.913	45223	97.687	96.700	98.070	97.246
Average Efficiency					97.302	96.586	98.070	97.632
Standard Deviation					1.833	2.384	1.380	1.447

TABLE 8

	Instance name	n	Avg. Deg.	Z-BC	Ordering_25	Ordering_25 nz	Ordering_50	Ordering_50 nz
1	GD06_theory	101	3.762	0	20.875	20.875	20.222	20.222
2	GD96_b	111	3.477	0	84.842	84.842	85.789	85.789
3	GD98_c	112	3.000	0	62.825	62.825	66.599	66.599